AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 109:281–293 (1999)

Estimation of Nonpaternity in the Mexican Population of Nuevo Leon: A Validation Study With Blood Group Markers

RICARDO M. CERDA-FLORES,[1–3] SARA A. BARTON,[2]
LUISA F. MARTY-GONZALEZ,[3] FERNANDO RIVAS,[4]
AND RANAJIT CHAKRABORTY[2]*
[1]*División de Genética, Centro de Investigación Biomédica del Noreste, Instituto Mexicano del Seguro Social, Administración de Correos 4, Monterrey, Nuevo León, 64720, Mexico*
[2]*Human Genetics Center, University of Texas School of Public Health, Houston, Texas 77225*
[3]*Facultad de Ciencias Biologicas, Universidad Autónoma de Nuevo León, Monterrey, 66450, Mexico*
[4]*División de Genética, Centro de Investigación Biomédica de Occidente, Instituto Mexicano de Seguro Social, Guadalajara, Jalisco, 44310 Mexico*

KEY WORDS Mexican population; blood groups; nonpaternity rate; exclusion probability; paternity index

ABSTRACT A method for estimating the general rate of nonpaternity in a population was validated using phenotype data on seven blood groups (A_1A_2BO, MNSs, Rh, Duffy, Lutheran, Kidd, and P) on 396 mother, child, and legal father trios from Nuevo León, Mexico. In all, 32 legal fathers were excluded as the possible father based on genetic exclusions at one or more loci (combined average exclusion probability of 0.694 for specific mother-child phenotype pairs). The maximum likelihood estimate of the general nonpaternity rate in the population was 0.118 ± 0.020. The nonpaternity rates in Nuevo León were also seen to be inversely related with the socioeconomic status of the families, i.e., the highest in the low and the lowest in the high socioeconomic class. We further argue that with the moderately low (69.4%) power of exclusion for these seven blood group systems, the traditional critical values of paternity index ($PI \geq 19$) were not good indicators of true paternity, since a considerable fraction (307/364) of nonexcluded legal fathers had a paternity index below 19 based on the seven markers. Implications of these results in the context of genetic-epidemiological studies as well as for detection of true fathers for child-support adjudications are discussed, implying the need to employ a battery of genetic markers (possibly DNA-based tests) that yield a higher power of exclusion. We conclude that even though DNA markers are more informative, the probabilistic approach developed here would still be needed to estimate the true rate of nonpaternity in a population or to evaluate the precision of detecting true fathers. Am J Phys Anthropol 109:281–293, 1999.

Use of genetic markers in determining whether or not a specified man is the biological father of a given child has a long history (Wiener et al., 1930; Essen-Möller, 1938), and the methods for determining how often an alleged father can be excluded as the biological father through genetic testing are well-known (reviewed in Neel and Schull, 1954; Chakraborty et al., 1974; Evett and Weir, 1998). While most empirical data on the efficiency of genetic testing for detecting nonpaternity come from cumulative studies of contested cases where paternity is in dispute, determination of the general rate of nonpaternity in a population is important

Grant sponsor: US National Institutes of Health; Grant numbers: GM 41399, GM 45861; Grant sponsor: U.S. National Institute of Justice; Grant numbers: 96-IJ-TX-0023, 98-LB-VX-K019.

*Correspondence to: Dr. Ranajit Chakraborty, Human Genetics Center, University of Texas School of Public Health, PO Box 20334, Houston, TX 77225. E-mail: rc@hgc9.sph.uth.tmc.edu

Received 14 April 1998; accepted 23 March. 1999.

for several reasons. For example, unrecognized nonpaternity adversely affects analysis of genetic data collected from a population. In linkage analysis, the estimate of recombination fraction is inflated in the presence of undetected nonpaternity; consequently, it reduces the power of detecting linkage. The sensitivity of estimating transmission probabilities of genetic disorders is also seriously affected in the presence of nonpaternity (Bonaiti-Pellié et al., 1992). Furthermore, the nonpaternity rate in a population, being at least partially dictated by cultural, behavioral, and social attributes of married couples, is a naturally justified topic of any physical anthropological investigation.

Using genetic data, several investigators estimated the proportion of children for whom the legal (or stated) father was not the biological father (Edwards, 1957; Ashton, 1980; Salmon et al., 1980; Ritz, 1985; Peñaloza et al., 1986; Macintyre and Sooman, 1991; Le Roux et al., 1992; Lincoln and Syndercombe, 1992; Sasse et al., 1994). While the rates vary across populations (from $<$1–30%), several features of such population-based determinations of rates of nonpaternity from genetic data are worth noting. The proportion of paternity exclusions detected through genotypic inconsistencies (incompatibilities from Mendelian rules) cannot be equated to the proportion of nonpaternity in the population, since these two rates can differ in either direction. Each genetic marker used singly, or in combination with others, always fails to exclude a certain fraction of nonpaternity. In constrast, genotypic mismatch, caused by typing errors, should not be ascribed to nonpaternity. Furthermore, although rare for traditional blood group markers, occasional mutations may mimic nonpaternity (Rothman et al., 1981; Chakraborty et al., 1996). Therefore, population-based data on genetic markers from stated family relationships have to be analyzed appropriately to estimate the nonpaternity rates taking these factors into account.

Sasse et al. (1994) suggested such a method where the incomplete detectability of nonpaternity through genetic markers was taken into account. In contrast, Rothman et al. (1981) and Chakraborty et al. (1996) discussed statistical methods of estimating mutation rates in the presence of nonpaternity in the population. Methods for detecting and correcting genotyping errors in databases are also discussed by others (e.g., Lathrop et al., 1983; Brzustowicz et al., 1993; Ehm et al., 1996). However, apart from suggesting how in individual cases this type of error can be inferred (Ashton, 1980; Chakraborty and Ryman, 1981), there is no general method for adjusting estimates of nonpaternity for errors of genotyping. Since genotyping errors in a laboratory are generally independent of actual nonpaternity, indirect procedures (such as the one described in Chakraborty and Schull, 1976) may be used to check whether any substantial fraction of detected genotyping incompatibilities are truly due to genotyping errors.

The purpose of this research was to use phenotype data on seven blood group systems in a sample of 396 (stated) mother-child-father trios from Monterrey, Mexico, to address some of these issues at a population level. In particular, we propose a method through which the empirical proportions of exclusions found in the genetic data can be adjusted, taking into account the incomplete power of excluding nonpaternity, so that the former can be used to estimate the rate of nonpaternity in the population. Consistency of the observed and expected distributions of the number of loci showing exclusions demonstrates that, in general, the exclusions are due to true nonpaternity, and they are not due to laboratory typing errors. Finally, we describe another utility of the estimated rate of nonpaternity in a population. Incorporating this estimate in the relationship between the paternity index (PI) and exclusion probability for each mother-child pair, we argue that the use of standard critical values for the paternity index (such as PI $\geq$ 19) for inferring paternity fails to identify all true fathers. We contend that this is mainly due to the low probability of exclusion as determined from the seven blood group markers, used for paternity testing in Mexico City (Armendares et al., 1985). Possible reasons for the high rates of nonpaternity (0.118 ± 0.020) estimated from the sample are also discussed. Although the data that we use (i.e., seven serological markers) to

validate our theoretical approach are somewhat old-fashioned, we argue that the theory is applicable for more recently introduced DNA markers as well, for which the combined exclusion probability can be much larger than the one depicted by the blood-group loci.

MATERIALS AND METHODS

Sample description and loci studied

Data analyzed in this research were collected from the Obstetrics and Gynecology Hospital in Monterrey, Nuevo León, Mexico in 1983. The sample was composed of 396 trios (mother, child, and stated father). Immediately following the birth of a newborn child in the hospital, the medical and laboratory personnel collected 5 ml of blood from the mother and child (from the umbilical cord) in tubes previously prepared with heparin. Blood samples from the fathers (with proper identification) were also collected on the same day (5 ml blood in heparin-prepared tubes). In addition to the blood samples, for each family a detailed questionnaire was administered to collect information on their socioeconomic status, geographic location of the birthplaces of the parents and of all four grandparents (by state and zone of the country), parental ages, length of marriage, birth information on previous children, and any congenital abnormality of the newborn. The socioeconomic status of each family was classified according to the criteria described in Bronfman et al. (1988). For simplicity, we used only three socioeconomic classes: low, medium, and high.

Appropriately labeled blood samples were processed at the Centro de Investigación Biomédica del Noreste laboratory, Instituto Mexicano del Seguro Social (IMSS), where laboratory personnel were not aware of the relationships of the samples analyzed. Phenotype determinations were made for each sample for the seven blood markers (A_1A_2BO, Rh (with -C, -c, -D, -E, and -e antisera), MNSs, Duffy (with Fya and Fyb), Lutheran, Kidd, and P) according to the procedure described in Crawford et al. (1970) and Lapinski et al. (1978). Identification required for each father, and independent collection of demographic information from the mothers and fathers, established the accuracy of the stated family relationship for each trio. All phenotypic discordances noted in the laboratory data were retyped to eliminate laboratory typing errors. At this point let us note that since the blood samples from these 396 families are no longer available, the blood group data that we used in the present analysis cannot be supplemented by more recent molecular data (e.g., microsatellite DNA markers) that can be implemented in similar studies.

Statistical methods

Phenotypic data on the seven blood group markers were stored in a computer for statistical analysis. Since no consanguinity was noted among the spouses, from 396 pairs of mothers and fathers, the observed phenotype distributions for each marker were used to calculate allele frequencies, using the maximum likelihood method (Reed and Schull, 1968). The homogeneity of allele frequency distributions in the sample of mothers and fathers was tested (by a contingency table analysis) for all markers before pooling the data from both parents.

From the observed phenotypes of all 396 trios, inconsistencies were searched to detect parentage exclusions, and the markers where discordances were found. In all cases the mother's phenotypes were compatible with the newborn's. For each mother-child phenotype, the expected chance of paternity exclusion was computed using the observed allele frequencies in the pooled sample of both parents, for which the computer routine MENDEL (Version 3.3; Lange et al., 1988) was used. The average probability of exclusion for each locus was computed using the respective formulae for each system (Wiener, 1952; Boyd, 1954; Rust, 1972).

Since the statistical features of the observed exclusions are dependent on each mother-child phenotype for each system, all subsequent analyses were based on paternity exclusion probabilities given the mother-child phenotype pair (MC) as observed in each case. We denote these probabilities as P_{ij}, for the i-th MC pair at the j-th locus (i = 1, 2, ..., n; j = 1, 2, ..., K). For the present work, the specific values were n = 396, and K = 7.

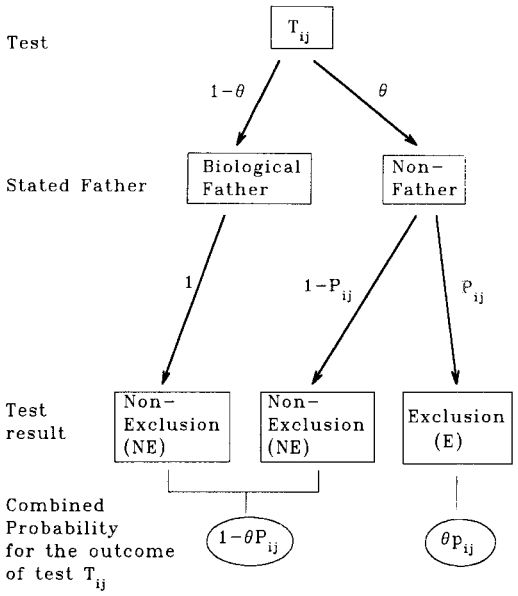

Fig. 1. Schematic design of estimation of nonpaternity in a population by genetic testing at several markers.

For estimating the rate of nonpaternity in the population, we applied the method of Sasse et al. (1994), using the complete data. We assume that this is the true rate of nonpaternity in the population. For each specific test (T_{ij}; i = 1, 2, ..., n; j = 1, 2, ..., K) for the i-th MC pair and j-th locus, the two possible outcomes of exclusion (E) and non-exclusion (NE) occur with the probabilities θP_{ij} and 1 - θP_{ij}, respectively, as shown in Figure 1. Denote the result for the test T_{ij} by an indicator variable δ_{ij}, defined by

$$\delta_{ij} = \begin{cases} 1, & \text{if } T_{ij} \text{ results in an exclusion (E),} \\ 0, & \text{if } T_{ij} \text{ results in a non-exclusion (NE),} \end{cases} \quad (1)$$

so that from Figure 1 we have the probability distribution given by

$$\Pr.(\delta_{ij} = 1) = \theta P_{ij}, \quad (2a)$$

and

$$\Pr.(\delta_{ij} = 0) = 1 - \theta P_{ij}, \quad (2b)$$

for i = 1, 2, ..., n; and j = 1, 2, ..., K.

Note that this formulation (see Fig. 1) clearly distinguishes the true nonpaternity rate (θ) in the population from the exclusionary power (P_{ij}) for the i-th MC pair at the j-th locus. For example, for a biallelic locus if the mother as well as the child are heterozygous, P_{ij} becomes zero irrespective of any nonzero value of θ in the population.

To proceed with the estimation of θ, first we assume that the nonpaternity rate (θ) is the same for all MC pairs. Since the test results of T_{ij} are mutually independent, the combined likelihood function (L) for the total sample is given by

$$L = \prod_{i=1}^{n} \prod_{j=1}^{K} [1 - \theta P_{ij}]^{1-\delta_{ij}} \cdot [\theta P_{ij}]^{\delta_{ij}}. \quad (3)$$

The maximum likelihood estimator of θ can now be obtained by maximizing L over θ (i.e., by solving the equation $\delta \log L / \delta\theta = 0$), which requires an iterative solution of the equation

$$\theta = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{K} \left\{ \frac{P_{ij}(1 - \delta_{ij})}{1 - \theta P_{ij}} \right\}}. \quad (4)$$

We used the scoring method as described by Rao (1965) to solve Equation (4). The sampling variance of the maximum likelihood estimate $\hat{\theta}$ is given by

$$V(\hat{\theta}) = \left[\sum_{i=1}^{n} \sum_{j=1}^{K} \left\{ \frac{P_{ij}^2}{(1 - \theta P_{ij})^2} + \delta_{ij} \right\} \right]^{-1}, \quad (5)$$

evaluated at $\theta = \hat{\theta}$.

Note that $\Sigma\Sigma\ \delta_{ij}$, summed over all MC pairs and all systems, represents the total number of exclusions observed in the sample of nK tests performed in the data (counting multiple loci exclusions for the same stated father separately).

Since the expected value of δ_{ij} is θP_{ij} (see Equations 2a,b), for all i and j, as an initial value of θ in solving Equation (4), we used the moment estimator

$$\tilde{\theta} = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{K} \delta_{ij} \right\} \Big/ \left\{ \sum_{i=1}^{n} \sum_{j=1}^{K} P_{ij} \right\}, \quad (6)$$

which by itself suggests that the observed number of exclusions in any empirical data from the population cannot be directly equated to estimate the nonpaternity rate,

because each $P_{ij} < 1$, and for some MC pairs the specific value for P_{ij} can even be zero (e.g., for a biallelic system when the mother and child are both heterozygous).

Two modifications of the method proposed are relevant at this point. First, since multilocus exclusions for a given trio are considered separately (as independent information) in the above formulation, one might argue that a random male can often be excluded on the basis of multiple markers, and hence, the estimate obtained from Equation (4) or (6) may be biased in an upward direction. This can be circumvented by using the MC-specific cumulative probability of exclusion

$$P_i = 1 - \prod_{j=1}^{K} (1 - P_{ij}), \tag{7}$$

instead of considering each P_{ij} separately, and by defining a corresponding indicator variable δ_i instead of δ_{ij}, so that $\delta_i = 1$ if at least one of the test results shows an exclusion for the i-th MC pair, and $\delta_i = 0$, otherwise. Equations (4–6) will hold with the simple modification that all summations are taken only over i (each MC pair), and test results are recorded for the combined battery of markers.

Second, the assumption that the chance of nonpaternity is constant over all MC pairs can be relaxed by the method discussed in Sasse et al. (1994). In this approach, we assume that θ_i is the chance of nonpaternity for the i-th MC pair (i = 1, 2, ..., n). Since for any single MC pair, some of the individual loci may be uninformative for exclusionary purpose (i.e., $P_{ij} = 0$ for one or more j; particularly for blood group markers such as the ones used in this analysis), the combined probability of exclusion for the MC pair (Eq. 7) is the relevant informative criterion for estimation of θ_i, and δ_i becomes the relevant observation (Sasse et al., 1994). With varying θ, the log likelihood function for the total sample becomes

$$\ln L = \sum_{i=1}^{n} [(1 - \delta_i) \cdot \ln (1 - \theta_i P_i) + \delta_i \cdot \ln (\theta_i) + \delta_i \cdot \ln (P_i)], \tag{8}$$

so that the maximum likelihood estimate of θ_i has the closed-form solution

$$\hat{\theta}_i = \delta_i / P_i = \begin{cases} 1/P_i, & \text{if the stated father is excluded for the i-th MC pair} \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the estimate of the average nonpaternity rate in the population is given by

$$\bar{\theta} = \Sigma(1/P_i)/n, \tag{9}$$

where the summation now extends over all MC pairs showing exclusion (on the basis of cumulative data over all loci) of the respective stated fathers. The asymptotic variance of this estimator is given by Sasse et al. (1994) as

$$V(\bar{\theta}) = \Sigma(1/P_i^2)/n^2, \tag{10}$$

in which the summation is again over all MC pairs, showing exclusions of the respective stated fathers.

The complete data on each system, however, provide further information on the reliability of test results. For example, Chakraborty and Schull (1976) derived the expected distribution of the number of loci exhibiting exclusions when the stated father was not the true biological father of a child. This distribution can be obtained, based on the MC-specific exclusion probability for each system. Let $Q_m^r(i)$ denote the probability that a random male is excluded on the basis of r of the first m loci typed for the i-th mother-child pair (m = 1, 2, ..., K; r = 0, 1, 2, ..., m). Following Chakraborty and Schull (1976), Q_m^r satisfies the recurrence relation

$$Q_m^r(i) = Q_{m-1}^r(i) \cdot (1 - P_{im}) + Q_{m-1}^{r-1}(i) \cdot P_{im}, \tag{11}$$

$$Q_m^0(i) = \prod_{j=1}^{m} (1 - P_{ij}), \tag{12}$$

and

$$Q_m^m(i) = \prod_{j=1}^{m} P_{im}. \tag{13}$$

Expression (11) can be recursively computed for any given data. Since P_{ij} values are

different for each specific MC pair, these expected distributions are to be computed separately for each MC pair. To contrast with the observed distribution of the number of loci exhibiting exclusions in a sample where a fraction of $(1 - \theta)$ stated fathers are true biological fathers, the expected frequencies of observing exclusions at r of the K loci typed are given by

$$N_K^r = \begin{cases} (1-\theta) + \theta \sum_{i=1}^{n} Q_K^0(i), & \text{for } r = 0 \\ \theta \sum_{i=1}^{n} Q_K^r(i), & \text{for } r = 1, 2, \ldots, K. \end{cases} \quad (14)$$

At this point it is worthwhile to note that this analytical formulation of deriving the expected distribution of the number of loci exhibiting exclusions avoids the use of Monte Carlo simulations, as suggested recently by Calafell et al. (1999).

Finally, for all trios which showed no exclusion based on all seven markers typed, the PI for the stated father was computed using the full phenotypic information of the mother, child, and stated father (Lee et al., 1983).

RESULTS

Table 1 shows the maximum likelihood estimates (computed by using the algorithm of Reed and Schull, 1968) of allele frequencies at the seven loci obtained from the sample of 396 mothers and fathers, separately, as well as from the pooled sample of 792 parents. Since at no locus the allele frequency difference between the two parental samples was significant (at the 5% level), the allele frequency estimates from the pooled data (Table 1) were used in all subsequent analyses.

Table 2 shows the phenotypes of mother, child, and stated father for 32 trios where the stated father was excluded on the basis of at least one locus typed. In all, 32 exclusionary cases were observed in 396 trios, giving an apparent rate of exclusion of 8.1%. However, the actual rate of nonpaternity was higher, since the seven markers together could not detect all nonpaternity entirely. Table 3 shows the locus-specific numbers (and proportions) of exclusions observed along with the expected chance of excluding a random male, based on allele frequencies in the sample. Two such expectations were computed. One, relevant to the observed data, was based on the specific MC-pair phenotypes observed (shown in the third column of Table 3), while the other was for the locus-average, applicable to a random MC-pair. Thus, the numbers shown in the third column are the average values over all 396 MC pairs, and their comparisons with the locus-average exhibit how well the data are representative of the allele frequencies in the pooled sample. The rank correlation between the expected probabilities of exclusions based on allele frequencies for the specific MC phenotypes and the locus-average is perfect (Spearman's $\rho = 1$; $P \approx 1.0$; Kendall, 1947), although for all markers except A_1A_2BO, even in the large sample of 396 MC pairs, there are differences between the MC-pair-specific and locus-average exclusion probabilities. This confirms the assertion that a validation of statistical proper-

TABLE 1. Maximum likelihood estimates of allele frequencies at seven blood-group loci in the sample of 396 pairs of mothers and fathers and in the pooled sample of 792 parents

Loci	Allele	Frequency		
		Mothers	Fathers	Pooled
A_1A_2BO	A_1	0.093	0.098	0.096
	A_2	0.043	0.034	0.038
	B	0.056	0.061	0.058
	O	0.808	0.807	0.808
Rh	CDE	0.047	0.036	0.041
	CDe	0.361	0.394	0.378
	cDE	0.202	0.236	0.221
	cDe	0.290	0.201	0.236
	CdE	0.000	0.000	0.000
	Cde	0.040	0.000	0.020
	cdE	0.024	0.009	0.014
	cde	0.036	0.124	0.090
MNSs	MS	0.229	0.256	0.243
	Ms	0.290	0.266	0.278
	NS	0.157	0.153	0.155
	Ns	0.324	0.325	0.324
Duffy	Fy(a)	0.561	0.541	0.551
	Fy(b)	0.307	0.314	0.311
	Fy	0.132	0.145	0.138
Lutheran	Lu^a	0.068	0.056	0.062
	Lu^b	0.932	0.944	0.938
Kidd	Jk(a)	0.313	0.268	0.290
	Jk(b) + Jk	0.687	0.732	0.710
P	P_1	0.627	0.631	0.629
	P_2 + p	0.373	0.369	0.371

TABLE 2. Phenotypes of mother, child, and stated father where the father is excluded

Family identification number	Locus	Phenotypes of		
		Mother	Child	Stated father
One-locus exclusions				
15	Kidd	A−	A+	A−
57	Kidd	A−	A+	A−
61	Rh	DCCee	DCcee	DCCee
79	A_1A_2BO	O	B	O
86	MNSs	NNSS	MNSS	MNss
89	A_1A_2BO	O	B	A_1
99	A_1A_2BO	A_1	B	O
117	MNSs	MNSS	MNSs	MNSS
132	Rh	DCcEe	DCcee	DccEE
144	MNSs	MNSS	MNSs	MNSS
147	Rh	DccEe	DccEe	DCCee
186	MNSs	MMSS	MNSS	MNss
198	MNSs	MNss	MNSs	MNss
229	A_1A_2BO	O	A_2	B
275	A_1A_2BO	O	B	O
280	Rh	DCcEe	dccee	DccEE
319	Rh	DCCee	DCcee	DCCee
348	A_1A_2BO	O	B	O
354	A_1A_2BO	A_1	B	O
Two-locus exclusions				
53	Duffy	A−B+	A+B+	A−B+
	Lutheran	A−	A+	A−
176	A_1A_2BO	O	A_1	O
	Duffy	A−B+	A+B+	A−B−
196	Rh	DccEe	Dccee	DCCee
	Kidd	A−	A+	A−
214	Rh	Dccee	Dccee	DCCee
	MNSs	MNss	MMSs	MNss
222	Rh	DccEe	DCcee	Dccee
	Kidd	A−	A+	A−
234	A_1A_2BO	A_2	A_2B	A_1
	Rh	DCcEe	DccEe	DCCee
255	Rh	DCcEe	DccEE	DCcee
	MNSs	MNss	MMSs	MMss
257	A_1A_2BO	O	A_2	O
	MNSs	NNss	MNSs	MNss
262	MNSs	MNSS	NNSs	MNSS
	Kidd	A−	A+	A−
263	Rh	dccee	Dccee	DCCee
	Lutheran	A−	A+	A−
271	Rh	DCCee	DCcEe	Dccee
	Duffy	A−B+	A+B+	A−B−
304	Rh	DCcEe	DCCee	DccEe
	MNSs	MNSs	MNSS	MNss
Three-locus exclusions				
298	Rh	Dccee	Dccee	DCCee
	MNSs	NNss	NNss	NNSS
	Kidd	A−	A+	A−

ties of exclusionary results in a population-based study should be based on MC-pair specific phenotype/genotype data, and not on the locus-average values based on allele frequencies alone (Chakraborty, 1995). Of course, the close agreement of the MC-specific (69.4%) and locus-average (70.5%) cumulative exclusion probabilities based on seven markers indicates that the MC pairs in the sample of 396 trios are a reasonable representation of a random sample from the population.

TABLE 3. Locus-specific and combined observed proportions and probabilities of paternity exclusion based on the allele frequencies in the sample[1]

Locus	Number (proportion) of exclusions observed	Exclusion probability based on	
		MC pair[2]	Locus-average[3]
A_1A_2BO	10 (0.025)	0.157	0.143
Rh	14 (0.035)	0.372	0.340
MNSs	11 (0.028)	0.247	0.333
Duffy	3 (0.008)	0.152	0.102
Lutheran	2 (0.005)	0.029	0.048
Kidd	6 (0.015)	0.057	0.073
P	0 (0.000)	0.010	0.012
Combined	32 (0.081)	0.694	0.705

[1] The allele frequencies used in these computations are for the pooled sample of all mothers and fathers (see text for citations for the formulae used for the computations).
[2] Averages over 396 mother-child (MC) pairs, based on expectations for the given MC phenotypic combinations.
[3] Expectations for each locus for a random MC pair.

The actual numbers (or proportions) of exclusions observed in the sample (second column of Table 3) are obviously much lower than the expected probabilities of exclusions (columns three or four). This is natural, because the data contained a substantial fraction of true fathers, which should show no exclusion based on any marker typed. Nevertheless, a regression line of observed and expected proportion of exclusions passing through the origin (i.e., $E(\delta_{ij}) = \theta P_{ij}$; see Fig. 1) is an indirect test of the justification that the exclusionary results observed are not caused by laboratory typing errors. Figure 2 shows the results of this analysis, resulting in a regression coefficient of 0.104, which is somewhat lower than the moment or maximum likelihood estimators of the nonpaternity rate in the population discussed below.

Table 4 shows the estimates of the nonpaternity rate (θ) based on methods of moment and maximum likelihood, with the assumption of constant θ for MC pairs (Eqs. 4–6), as well as the one based on varying θ (Eqs. 9, 10). Numerically, under the assumption of constant θ, both methods yield very similar estimates, and the multiple locus exclusions of some of the stated fathers (13 of the 32 exclusions are found at 2 or 3 loci) do not appear to change the estimate of θ considerably. However, use of cumulative exclusion-

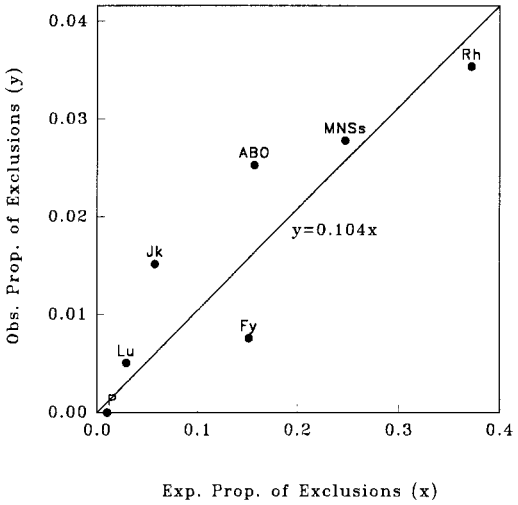

Fig. 2. Observed proportion of stated fathers excluded based on each locus vs. the expected probability of excluding a random male for the Mexican mother-child pairs from Nuevo León.

TABLE 4. Estimates of nonpaternity rate (θ) in Mexicans of Nuevo León, Mexico

Data and estimation method	Based on Individual loci	Based on All loci
Number of exclusions	46[1]	32
Estimates of nonpaternity		
Method of moments	0.113	0.117
Maximum likelihood		
Constant θ	0.114 ± 0.016	0.118 ± 0.020
Varying θ		0.101 ± 0.018

[1] Total number of exclusions counting each individual locus-specific exclusion separately (n = 396).

ary data over the seven loci makes the standard error of the maximum likelihood estimate somewhat larger than that obtained from data on locus-specific exclusions. When θ is assumed to vary over the MC pairs, cumulative data over all loci gave a somewhat lower estimate of the average θ (10.1%).

The distribution of the number of loci exhibiting exclusions for the entire data and its expectation based on the above estimates of nonpaternity rate in the population are shown in Table 5. The observed distribution is derived from data shown in Table 2, and the expected distributions were computed from Equation (14). In order to compare the effect of the assumption of constant vs.

TABLE 5. Observed and expected distributions of the number of loci exhibiting exclusions of stated fathers in the Mexican population of Nuevo León

No. of loci exhibiting exclusions	Frequencies		
		Expected based on	
	Observed	$\hat{\theta} = 0.118$	$\bar{\theta} = 0.101$
0	364	363.69	368.42
1	19	19.67	16.79
2	12	9.95	8.50
3	1	2.39	2.03
4	0	0.28	0.24
5	0	0.02	0.02
≥6	0	0.00	0.00
Total	396	396.00	396.00

varying θ over the MC pairs, in these computations we used the maximum likelihood estimates of θ (or average θ), using the cumulative exclusionary data over all loci (i.e., estimates used are $\hat{\theta} = 0.118$, and $\bar{\theta} = 0.101$). Comparison of the observed and expected distributions shows that the observed number of loci showing exclusions fits either of the two expected distributions adequately (goodness-of-fit $\chi^2 = 0.01$ for constant θ, and 0.80 for varying θ, with exclusions based on two or more loci classes merged), although the fit with the constant θ model appears slightly better. Figure 3 shows the comparison of observed and expected (constant θ) distributions graphically. The closeness of these distributions infers that the multiple-locus exclusions as well as the single-locus exclusions are not likely to have been caused by laboratory typing errors.

These computations indicate that in the present data, 32 of the 396 stated fathers were excluded based on at least one of the seven markers typed, and none of these were probably due to laboratory typing errors. In addition, on the basis of estimated θ, there could be an additional 8–15 (= 396θ - 32) true nonfathers who remained unexcluded among the stated fathers. In order to examine whether they could have been inferred from the PI values (for computational method of PI, see Lee et al., 1983), in Figure 4 we plotted the PI against the cumulative P_i for each MC pair for all nonexcluded (364) stated fathers. Traditionally, with a prior probability of 50%, PI ≥ 19 corresponds to a 95% probability of paternity. However, in the present data, of the 364 PI values, only 57 (= 15.7%) were higher than 19. In con-

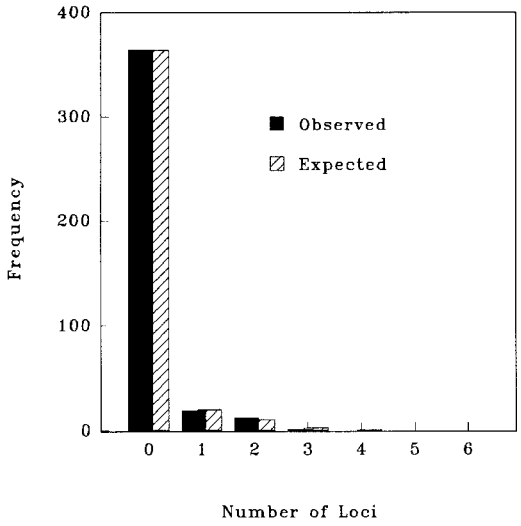

Fig. 3. Observed (solid bars) and expected (hatched bars) distributions of the number of loci exhibiting exclusions of stated fathers in the Mexican population of Nuevo León. The expected distribution is based on the maximum likelihood estimate of nonpaternity rate (θ), assuming constant value of θ for all mother-child pairs.

trast, with the estimated 11.8% nonpaternity, we expected approximately 350 true biological fathers in the sample. We ascribe this discrepancy to the low (69.4%) power of exclusion obtained for the MC pairs based on the seven markers typed.

DISCUSSION AND CONCLUSIONS

As mentioned before, Peñaloza et al. (1986) estimated the nonpaternity rate in a Mexican population from a sample collected from Mexico City. Their estimate (2.3%) was about 5-fold smaller than the one obtained here (11.8%). Peñaloza et al. (1986) did not adjust their estimate for the incomplete power of detecting nonpaternity by the eight markers they studied (ABO, Rh, MN, Ss, Duffy, P, acid phosphatase, Gc, and haptoglobin; with a reported combined power of exclusion of 80.3%). However, even this does not account for the difference of their estimate from ours (with a combined P_i for a random MC pair, an approximate method of moment estimate from their data would have been 2.9%, still substantially lower than ours). This adjustment may not be accurate for their data on 217 trios, because the published data of Peñaloza et al. (1986) did not provide an estimate of the combined P_i for the MC-specific phenotype pairs observed in their sample. Furthermore, their reported combined power of paternity exclusion for a random MC pair in the data from Mexico City treated the (C, c) and (D, d) alleles at the Rh locus, and (M, N) and (S, s) alleles at the MNSs locus as independent, which would have the effect of overestimating the combined power of exclusion.

Nevertheless, we contend that there may be a true variation of nonpaternity rates in different populations of Mexico. Several possible reasons for interpopulation differences of nonpaternity rates may be postulated.

First, in the published literature there are indications that the rate of nonpaternity is dependent on cultural, socioeconomic, and psychological factors. On the basis of data on seven blood group loci from 367 Caucasian and 96 African-American families, ascertained through children who were admitted as patients in the Children's Hospital in Detroit, or at the University of Michigan Hospital, Schacht and Gershowitz (1963) estimated that nonpaternity occurred at a rate of 10.1% in the African-American and 1.5% in the Caucasian children. While the Caucasian rate is comparable with that in the study of Brock and Shrimpton (1991) from families in England, other Caucasian population-based studies yielded varied rates of nonpaternity. For example, Sasse et al. (1994) obtained nonpaternity rates lower than 1% in Switzerland; Edwards (1957) estimated a 3.7% nonpaternity in West London Caucasians, but Ritz (1985) reported a nonpaternity rate of almost 10% in Caucasians from Munich and Copenhagen. Thus, in view of such wide variation, the difference between the present estimate of 10.1–11.8% from Nuevo León and 2.9% from Mexico City (Peñaloza et al., 1986) is not surprising.

Second, different surveys estimating nonpaternity rates used genetic markers with variable degrees of efficiency for detecting true nonpaternity. The theory presented above suggests that for the most appropriate adjustment, we should use the precise MC-pair genotype (phenotype) data for all trios. While this suggestion had been made earlier by Lincoln and Syndercombe (1992), from the published data, it is impossible to deter-

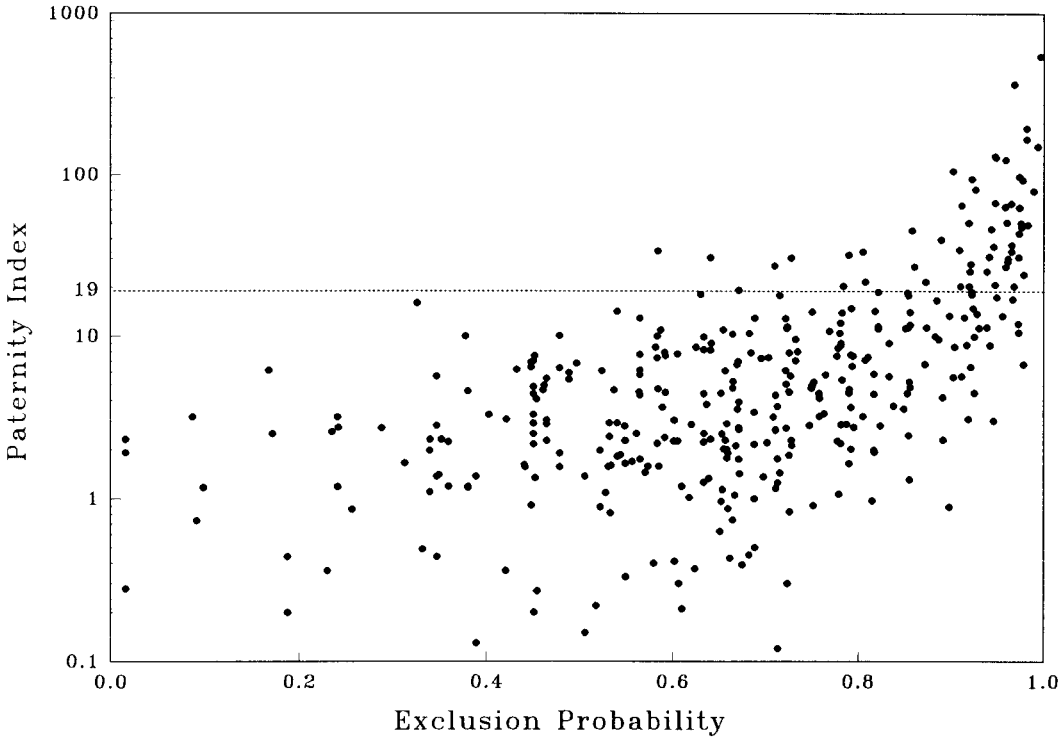

Fig. 4. Natural logarithm of paternity index (ln PI) of nonexcluded stated fathers plotted against the probability of exclusion for each mother-child pair. Of the 364 nonexcluded stated fathers, PI exceeded 19 for only 57 stated fathers.

mine the extent to which interpopulation variation is caused by inherent limitations of genetic markers. The theory presented here should provide guidelines for a uniform method of adjustment, applicable for any specific set of genetic markers used (including more efficient DNA markers).

Third, the interpopulation variation may in part simply indicate a sociological aspect of nonpaternity. Schacht and Gershowitz (1963) showed that, in their data, the nonpaternity rate was largest for the first-born child. Thus, they concluded that the first child was born more often out of wedlock than subsequent children within the families. In order to check this possibility in our sample, we grouped the observed exclusions by birth order of the children to examine if the observed rates of paternity exclusions differed by birth order. Figure 5 shows the data, reflecting that the exclusion rate by birth order varies from 5.71–11.53%, which is not significantly different from that of the total sample (8.08%, $\chi^2_6 = 2.63$, $P > 0.83$). Thus, we conclude that there is no birth-order effect on the nonpaternity rate in the population of Nuevo León. The birthplaces of the parents (i.e., both from Nuevo León, one outside Nuevo León, or both outside Nuevo León) had no effect on the nonpaternity rate, and neither did the age of parents (data not shown). The most likely explanation for the high rate of nonpaternity in Nuevo León in comparison to the Mexican population studied by Peñaloza et al. (1986) is probably the low socioeconomic status of families in Nuevo León, as was the situation found in the study of Schacht and Gershowitz (1963).

To check this assertion further, we subdivided our data by socioeconomic classes of parents that yielded 186 families in low, 157 in medium, and 53 in high socioeconomic class. These subdivisions exhibited 26, 5, and 1 fathers excluded from paternity based on at least one locus exclusion, respectively.

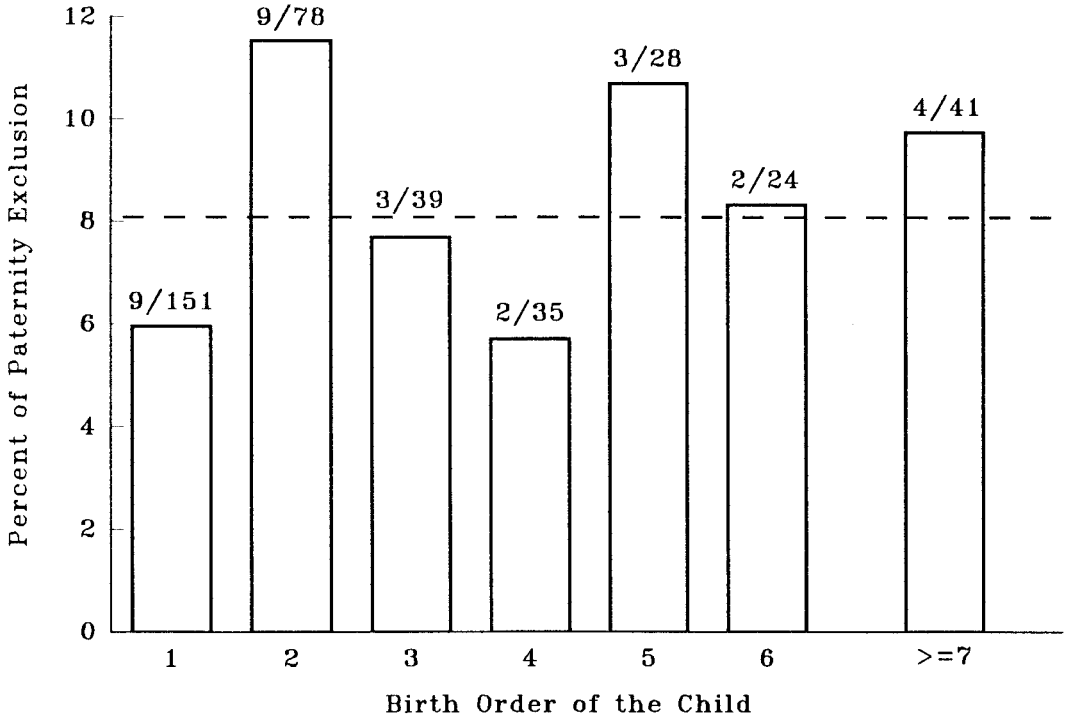

Fig. 5. Rate (in percent) of paternity exclusion in children, by birth order of children, in Nuevo León, México. Dotted line represents rate of paternity exclusion in the pooled data.

The maximum-likelihood estimates of nonpaternity rates based on the constant θ model for the combined systems (as data were sparse for individual locus computations) for these three groups were 0.198 ± 0.036, 0.047 ± 0.021, and 0.029 ± 0.029, respectively, consistent with the hypothesis that the nonpaternity rate is inversely related with socioeconomic status of the families in our samples as well.

Variation of nonpaternity rates in populations classified by socioeconomic status was also reported by Baker and Bellis (1995), who stated rates such as 1% in high socioeconomic populations of the United States and Switzerland, 5–6% in medium-income groups of the United Sates and England, and 10–30% in low socioeconomic groups of England, France, and the United States.

The present validation demonstrated that blood-group loci, in spite of their limited efficiency (about 71%) in detecting nonpaternity, can be used to show that genetic markers can reasonably estimate the true nonpaternity at a population level. However, some implications of our results in the context of selection of genetic markers to conduct paternity testing are obvious. For example, the seven blood-group loci analyzed here, that are commonly used in Mexico in conjunction with acid phosphatase, GC, and haptoglobin (see Armendares et al., 1985; Peñaloza et al., 1986), are not adequate to detect true biological fathers. In the present sample, a substantial fraction of the true fathers had a paternity index much lower than 19 with the seven markers studied. This, we contend, was primarily due to the low power of exclusion provided by a considerable proportion of MC-pairs. This conclusion was reached from the observed distribution of the combined power of exclusion for 396 MC pairs, and the distribution of PI (based on all seven markers) for 364 nonexcluded stated fathers, as graphically shown in Figure 4. Thus, in a society where the nonpaternity rate is as high as 11.8%, in order to exclude most of the nonfathers, as well as to detect

almost all true fathers, a more efficient battery of genetic markers should be used. Such markers are available, and can easily be implemented with DNA typing (Pena and Chakraborty, 1994; Alford et al., 1994). Even though paternity testing by DNA markers is more expensive than simple serological typing on a per case basis, in the long run, by identifying the true fathers more often, the cost of child support by public funding can be reduced. Furthermore, use of population-based family data for genetic analysis becomes more reliable when the assessment of paternity is done with an efficient selection of markers.

However, we reiterate that even if more informative DNA markers are used, the theory discussed here will still be necessary to estimate the nonpaternity rate in the population, although the difference between the empirically observed rate and the one estimated (higher than the observed one) may be smaller. This is so, because the combined power of exclusion for DNA markers is generally much higher than that of blood groups, but still, in a small proportion of cases, wrongly alleged fathers may be unexcluded. The necessity of the probabilistic approach will be accentuated with the use of DNA markers even further, since apparent exclusion based on one DNA marker can be due to mutations that are far more probable at minisatellite or microsatellite loci than in blood groups (Weber and Wong, 1993; Chakraborty et al., 1997). Elsewhere, we showed that there is a high probability ($>2\%$) of one-locus exclusion that can be caused by mutations when most informative DNA markers are employed in paternity testing (Chakraborty and Stivers, 1996).

Finally, our estimate of nonpaternity rate (11.8%) in the Mexican population of Nuevo León also has implications beyond sociogenetic considerations. For example, since we contend that the detected paternity exclusions in the present study were not influenced by laboratory errors, then should this rate of nonpaternity apply to the population at-large in Nuevo León, caution has to be exercised in interpreting the genetic epidemiological data from such a population. Undetected nonpaternity in family data adversely affects detection of familial resemblance of risk-factor variables (Chakraborty et al., 1980). It induces systematic negative biases in estimates of the genetic component of variation of quantitative traits (Lathrop, 1980), and leads to wrong conclusions regarding modes of inheritance of traits in segregation analysis (Bonaiti-Pellié et al., 1992). Further, undetected nonpaternity tends to inflate the estimate of the recombination fraction (Ehm et al., 1996) in linkage analysis. However, when family data cannot be directly checked for nonpaternity, our estimate may be used to make appropriate adjustments for possible nonpaternity when the family data are collected from this population.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance and contributions of the physicians, nurses, and medical technologists of the Hospital de Ginecología y Obstetricia de Monterrey, Instituto Mexicano del Seguro Social, Nuevo León, Mexico.

LITERATURE CITED

Alford RL, Hammond HA, Coto I, Caskey CT. 1994. Rapid and efficient resolution of parentage by amplification of short tandem repeats. Am J Hum Genet 55:190–195.

Armendares S, Lisker R, Pérez Briceño R, Buentello L. 1985. Sensibilidad de seis marcadores geneticos para la identificacion de ilegitimidad en la Ciudad de Mexico. Rev Invest Clin 37:27–30.

Ashton GC. 1980. Mismatches in genetic markers in a large family study. Am J Hum Genet 32:601–613.

Baker RB, Bellis MA. 1995. Human sperm competition: copulation, masturbation, infidelity. New York: Chapman and Hall. p 60, 139.

Bonaiti-Pellié C, Poisson N, Bechtel Y, Bechtel P. 1992. Sensitivity of transmission probabilities to paternity exclusion in segregation analysis. Genet Epidemiol 9:67–71.

Boyd WC. 1954. Tables and nomograms for calculating chances of excluding paternity. Am J Hum Genet 6:426–433.

Brock DJH, Shrimpton AE. 1991. Nonpaternity and prenatal genetic screening. Lancet 338:1151.

Bronfman M, Guisacafre H, Castro V, Castro R, Gutierrez G. 1988. Strategies for improving the therapeutic patterns used in acute diarrhea in primary medical care units. II. The measurement of inequality: a methodologic strategy, analysis of the socioeconomic features of the sample. Arch Invest Med (Mex) 19:351–360.

Brzustowicz LM, Mérette C, Xie X, Townsend L, Gilliam TC, Ott J. 1993. Molecular and statistical approaches to the detection and correction of errors in genotype databases. Am J Hum Genet 53:1137–1145.

Calafell F, Shuster A, Speed WC, Kidd JR, Black FL, Kidd KK. 1999. Genealogy reconstruction from short tandem repeat genotypes in an Amazonian population. Am J Phys Anthropol 108:137–146.

Chakraborty R. 1995. Exclusion of paternity by DNA-PCR technology: comment on validation studies. In: Proceedings from the Fifth International Symposium on Human Identification 1994. Madison, WI: Promega. p 19–22.

Chakraborty R, Ryman N. 1981. Use of odds of paternity computations in determining the reliability of single exclusions in paternity testing. Hum Hered 31:363–369.

Chakraborty R, Schull WJ. 1976. A note on the distribution of the number of exclusions to be expected in paternity testing. Am J Hum Genet 28:615–618.

Chakraborty R, Stivers DN. 1996. Paternity exclusion by DNA markers: effects of paternal mutations. J Forensic Sci 41:671–677.

Chakraborty R, Shaw M, Schull WJ. 1974. Exclusion of paternity: The current state of the art. Am J Hum Genet 26:477–488.

Chakraborty R, Weiss KM, Schull WJ. 1980. A test of randomness of the occurrences of a disease trait in familial or other similar ordered sequences of epidemiological data. Proc Natl Acad Sci USA 77:2974–2978.

Chakraborty R, Stivers DN, Zhong Y. 1996. Estimation of mutation rates from parentage exclusion data: application to STR and VNTR loci. Mutat Res 354: 41–48.

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci USA 94:1041–1046.

Crawford MN, Gottman FE, Gottman CA. 1970. Microplate system for routine use in blood bank laboratories. Transfusion 10:258–263.

Edwards JH. 1957. A critical examination of the reputed primary influence of ABO phenotype on fertility and sex ratio. Br J Prev Soc Med 11:79–89.

Ehm MG, Kimmel M, Cottingham RW Jr. 1996. Error detection for genetic data using likelihood methods. Am J Hum Genet 58:225–234.

Essen-Möller E. 1938. Die beweisbraft der Ahnlichkeit im Vaterschaftsnachweis, theoretische Grundlagen. Mitt Anthropol Ges Wien 68:9–53.

Evett IW, Weir BS. 1998. Interpreting DNA Evidence. Sunderland, MA: Sinauer Associates, Inc. p 163–177.

Kendall MG. 1947. The theory of statistics, volume 1. London: Charles Griffin and Co. p 380–390.

Lange K, Weeks DE, Boehnke M. 1988. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol 5:471–172.

Lapinski FJ, Crowley KM, Merrit CA, Henry JB. 1978. Use of microplate methods in paternity testing. Am J Clin Pathol 70:766–769.

Lathrop GM. 1980. Statistical methods in pedigree analysis and the reliability of pedigree data. Ph.D. dissertation. Seattle: University of Washington.

Lathrop GM, Hooper AB, Huntsman JW, Ward RH. 1983. Evaluating pedigree data: I. The estimation of pedigree error in the presence of marker mistyping. Am J Hum Genet 35:241–262.

Lee CL, Allen FH, Lebeck LK. 1983. Estimation of paternity index. In Walker RH, editor. Inclusion probabilities in parentage testing. Arlington, VA: American Association of Blood Banks. p 89–102.

Le Roux MG, Pascal O, Andre MT, Herbert O, David A, Moisan JP. 1992. Non-paternity and genetic counseling. Lancet 340:607.

Lincoln PJ, Syndercombe D. 1992. Rates of nonpaternity. Lancet 340:1108.

Macintyre S, Sooman A. 1991. Non-paternity and prenatal screening. Lancet 338:869–871.

Neel JV, Schull WJ. l954. Human heredity. Chicago: University of Chicago Press. p 318–330.

Pena SDJ, Chakraborty R. 1994. Paternity testing in the DNA era. Trends Genet 10:204–209.

Peñaloza R, Nuñez C, Alatorre S, Lagunes R, Garcia Escobar B, Salamanca F, Zavala C. 1986. Frecuencia de paternidad extraconyugal en una muestra de la poblacion Mexicana. Rev Invest Clin 38:287–291.

Rao CR. 1965. Linear statistical inference and its applications. New York: Wiley and Sons. p 302–305.

Reed TE, Schull WJ. 1968. A general maximum likelihood estimation program. Am J Hum Genet 20:579–580.

Ritz E. 1985. The clinical spectrum of hereditary nephritis. Nephrol Forum Kidney Int 27:83.

Rothman ED, Neel JV, Hoppe FM. 1981. Assigning a probability of paternity in apparent cases of mutation. Am J Hum Genet 33:617–628.

Rust PF. 1972. On the probability of detecting nonpaternity through the Rh blood-group system. Am J Hum Genet 24:54–57.

Salmon D, Seger J, Salmon C. 1980. Expected and observed proportion of subjects excluded from paternity by blood phenotypes of a child and its mother in a sample of 171 families. Am J Hum Genet 32:432–444.

Sasse G, Mller H, Chakraborty R, Ott J. 1994. Estimating the frequency of nonpaternity in Switzerland. Hum Hered 44:337–343.

Schacht LE, Gershowitz H. 1963. Frequency of extramarital children as determined by blood groups. In: Proceedings of the Second International Congress of Human Genetics, volume II. Rome: Instituto G. Mendel. p 894–897.

Weber JL, Wong C. 1993. Mutation of human short tandem repeats. Hum Mol Genet 8:1123–1128.

Wiener AS. 1952. Heredity of the MNS blood types. Am J Hum Genet 4:37–53.

Wiener AS, Lederer M, Polayes SH. 1930. Chances of proving non-paternity, considering multiple allelomorphic genes. J Immunol 19:259–281.